

Norbert KNOCHE, Essen & Detlef LIND, Wuppertal

Strukturanalysen zur Mathematikleistung und eine differentielle Itemanalyse der PISA-2000-Items zu den Faktoren Bildungsgang und Geschlecht

1. Korrelation und multivariate Regression

Als auffälliger Befund der PISA-2000-Studie wurden bereits in ARTELT, C., DEMMRICH, A., BAUMERT, J. (2001), KNOCHE, N., LIND, D. et al. (2002) und KNOCHE, N., LIND, D. (2004a) die hohen Korrelationen zwischen den Leistungen in den Tests zu *Lesekompetenz*, *Mathematische Grundbildung* und *Naturwissenschaftliche Grundbildung* berichtet. Die auf der 35584 Schüler(innen) umfassenden Datenbasis der nationalen PISA-Stichprobe aller 15-Jährigen berechneten Korrelationskoeffizienten liegen z. B. alle nahe bei 0,8. Ob sich in ihnen eher gemeinsame Hintergrundvariablen und weniger „direkte Abhängigkeiten“ manifestieren, kann man mit Hilfe von *Pfadmodellen* untersuchen:

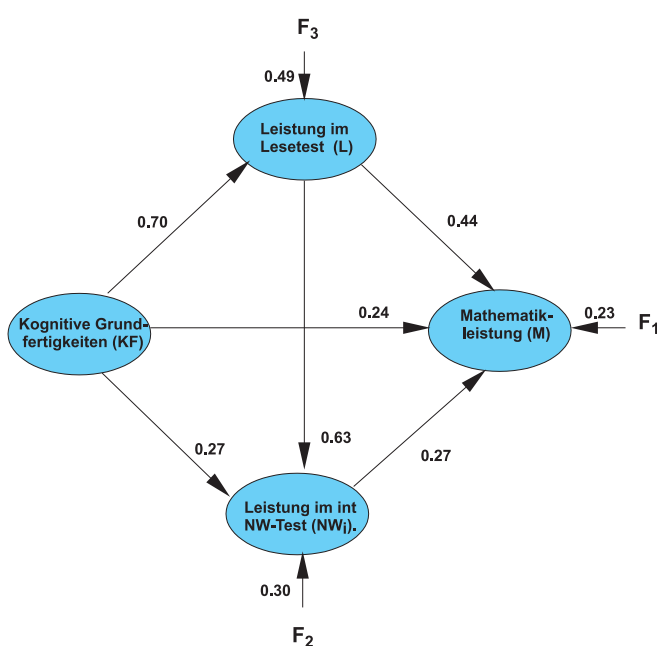


Abbildung 1:

Pfaddiagramm zur Beurteilung des Effektes der Variablen

Kognitive Grundfertigkeiten, Leseleistung und Leistung im internationalen Naturwissenschaftstest auf die Mathematikleistung.

Alle Variablen sind standardisiert.

Ein Pfadmodell ist ein verallgemeinertes multivariates *Regressionsmodell*, auch *Lineares Strukturgleichungsmodell* genannt, bei dem eine Variablenmenge V gegeben ist und ausgewählte Variablen aus V durch Linearkombinationen von „Prediktorvariablen“ aus V im Sinn der Regressionsrechnung approximiert werden. Dabei setzt man jede „Zielvariable“ Y in der Form $Y_i = a_{i0} + \sum a_{ik} \cdot X_k + F_i$ an und verlangt, dass die Varianz der „Fehlervariablen“ F minimiert wird.

Das zu Abbildung 1 gehörende Modell hat die einfache Gestalt

$$\begin{aligned} Y_1 &= a_{12} Y_2 + a_{13} Y_3 + c_{11} X + F_1 \\ Y_2 &= a_{23} Y_3 + c_{21} X + F_2 \\ Y_3 &= c_{31} X + F_3 \end{aligned}$$

mit $X = KF$, $Y_1 = M$, $Y_2 = NW_i$, $Y_3 = L$.

Darin bezeichnen die Variablen F_i die in Regressionsmodellen auftretenden Fehlervariablen, für die $E(F_i) = 0$ angenommen wird.

Unter den Voraussetzungen $\text{cov}(F_i, X_j) = 0$, $\text{cov}(F_i, F_j) = 0$ und der hier erfüllten Forderung, dass die Kovarianzmatrix der Daten positiv definit ist, liefert der Datensatz nach *Standardisierung* aller Variablen die in dem Pfaddiagramm an den Pfeilen eingetragenen Regressionskoeffizienten. Die an den Pfeilen von F_1 , F_2 und F_3 stehenden Zahlen τ_1 , τ_2 und τ_3 geben die Varianz der jeweiligen Fehlervariablen F_1 , F_2 und F_3 an.

Bei der Interpretation solcher Modelle muss man sich vergegenwärtigen, dass die Modellstruktur durch den Beobachter *a priori* vorgegeben ist und eine Bewertung der Bedeutung einer Variablen für eine andere nur im Rahmen der gemachten Vorgaben möglich ist.

Es hat sich eingebürgert, die Größe von Regressionskoeffizienten als Maß für den „Einfluss“ der Prediktorvariablen zu bezeichnen. Diese Größe (auch *Effekt* genannt) ist aber ein Maß für die Bedeutung der Prediktorvariablen in der *linearen Regression* und darf nicht im Sinne von kausaler Abhängigkeit von Werten verstanden werden. Das Modell in Abbildung 1 beschreibt eine Momentaufnahme nach *abgeschlossenen* Lernprozessen, womit sich stark vergrößerte Deutungen wie z. B. „erhöht man die Lesekompetenz von Schülern (durch unterrichtliche Maßnahmen) um eine Standardabweichung, dann erhöht sich die Mathematikleistung um etwa eine halbe Standardabweichung“ verbieten.

Bei der Beurteilung des Effektes einer Variablen auf eine andere Variable in einem Pfadmodell unterscheidet man zwischen „direkten“ und „indirekten“ Effekten. So ist der direkte Effekt von KF auf NW_i gleich 0.27, der über L vermittelte indirekte Effekt gleich $0.70 \cdot 0.63 = 0.44$. Der „totale“ Effekt von KF auf NW_i ist dann die Summe des direkten und indirekten Effekts. Entsprechend ergibt sich der totale Effekt von KF auf M zu 0.75 und der totale Effekt von L auf M ergibt sich zu 0.61.

2. Ein Pfadmodell zur Beschreibung der Mathematikleistung

Wir wollen nun explorativ der Bedeutung der Faktoren *Kognitive Grundfertigkeiten*, *Sozioökonomischer Status*, *Interesse an Mathematik* und *Selbstkonzept Mathematik* für die *Leistung im Mathematiktest* nachgehen.

Dabei beziehen wir wieder die Variablen L und NW_i in das Pfadmodell ein und stellen zwei Pfadmodelle vor, die mit denselben postulierten Pfaden getrennt auf den Teilstichproben der 15jährigen Jungen und Mädchen berechnet wurden.

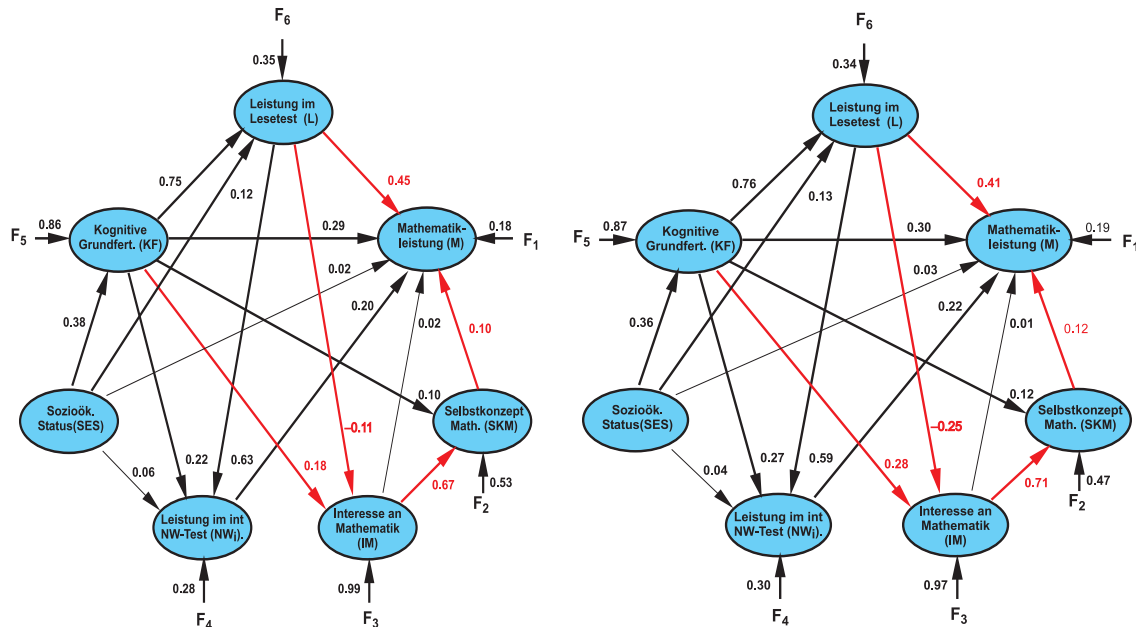


Abbildung 2: Pfaddiagramme für die wechselseitigen Beziehungen der Variablen *Kognitive Grundfertigkeiten* (KF), *Interesse an Mathematik* (IM), *Selbstkonzept Mathematik* (SKM), *Sozioökonomischer Status* (SES), *Leseleistung* (L), *Leistung im internationalen Naturwissenschaftstest* (NW_i) und *Mathematikleistung* (M); links für die Jungen, rechts für die Mädchen.

Es geht dabei nur um *strukturelle* Aussagen, und nicht um die Diskussion von *geschlechtsspezifischen Leistungsunterschieden*.

Die aufgeklärten Varianzanteile stimmen in beiden Teilpopulationen außer beim *Selbstkonzept Mathematik* nahezu überein, insbesondere wird jeweils 82% bzw. 81% der Varianz der Mathematikleistung aufgeklärt. Es fällt auf, dass alle bedeutsamen Unterschiede von Pfadkoeffizienten mit der Variablengruppe KF – L – IM – SKM assoziiert sind (die Pfade sind in Abb. 2 rot gekennzeichnet).

Der direkte Effekt von KF auf IM im Rahmen des Modells ist mit 0.28 bei den Mädchen deutlich höher als bei den Jungen (nur 0.18) und der „Konflikt“ zwischen L und IM ist bei den Mädchen ebenfalls stärker als bei den

Jungen ausgeprägt (dies lässt sich aus den Pfadkoeffizienten -0.11 und -0.25 schließen).

Eine mögliche Ursache für diesen Befund liegt in den schon in STANAT, P. & KUNTER, M. (2001) diskutierten Unterschieden in den Lesegewohnheiten von Jungen und Mädchen und den daraus resultierenden Unterschieden in der Leseleistung. Unterstellt man, dass großes Interesse am Lesen und häufige Leseaktivitäten in Konkurrenz zur Beschäftigung mit Mathematik stehen können, so sollte sich dies in einem negativen Pfadkoeffizienten von L auf IM abbilden. Wegen der stärker ausgeprägten Leseneigung der Mädchen ist dieser Pfadkoeffizient dem Betrag nach über doppelt so groß wie der entsprechende Koeffizient bei den Jungen. Da die unmittelbaren Effekte von KF auf alle Variablen außer IM bei beiden Geschlechtern nahezu gleich sind, muss der Pfadkoeffizient von KF auf IM bei den Mädchen größer als bei den Jungen sein. Eine „griffige“ Formulierung der Art „bei Mädchen wirkt sich die Intelligenz stärker direkt auf das Mathematikinteresse als bei Jungen aus“ wird diesem komplexen Sachverhalt ebenso wenig gerecht wie „bei Jungen ist das Mathematikinteresse weniger direkt von der Intelligenz abhängig“ (!).

Auf diesem Hintergrund muss auch die Diskussion um das *mathematische Selbstkonzept* SKM gesehen werden. Der totale Effekt von *Kognitive Grundfertigkeiten* (KF) auf SKM ist im Rahmen des Modells mit $0,22$ für die Jungen deutlich kleiner als mit $0,32$ für die Mädchen. Damit einher geht der deutliche Unterschied des totalen Effekts der Leseleistung L auf SKM zwischen $-0,07$ für die Jungen und $-0,18$ für die Mädchen. Dies spricht noch einmal für die im Rahmen des *Mathematikinteresses* angesprochene Deutung, dass bei Schülerinnen Lesen und Mathematik häufiger zueinander in Konkurrenz stehen als bei Jungen.

3. Zum Begriff DIF

Bei Schwierigkeitsanalysen muss man zwischen Personenmerkmalen und Aufgabenmerkmalen unterscheiden. *Personenmerkmale* sind unter anderem das Geschlecht, das Alter, der besuchte Bildungsgang, aber auch Eigenschaften wie das Interesse an Mathematik usw. Zu den *Aufgabenmerkmalen* gehören unter anderem Anforderungskategorien bei der Bearbeitung der Aufgaben, Itemformate, der Kontext der Aufgabenstellung, usw.

Wir beschäftigen uns im folgenden zunächst mit Personenmerkmalen. Wir betrachten die Aufgaben auf den Teilpopulationen Bildungsgang und Geschlecht. Bei der Analyse unterschiedlichen Verhaltens dieser Teilpopulationen auf den Aufgaben kommen dann in einem zweiten Schritt auch Aufgabenmerkmale ins Spiel.

Zerlegt man eine Population in Teilpopulationen, so erlaubt CONQUEST die simultane Schätzung von Raschmodellen für die Teilpopulationen und die Schätzung eines Gesamtmodells für die gesamte Population. Von den geschätzten Parametern interessieren uns i. f. vorrangig die Itemschwierigkeiten in der Gesamtpopulation und den Teilpopulationen, am Rande auch die durchschnittlichen Leistungsfähigkeiten der einzelnen Populationen.

Im Titel des Vortrags taucht das Wort DIF auf. Als DIF eines Items bzgl. einer Teilpopulation bezeichnet man die Differenz zwischen der Itemschwierigkeit dieses Items in der Teilpopulation und der Gesamtpopulation bzw. die Differenz zwischen der Itemschwierigkeit des Items in der einen und der anderen Teilpopulation. Trägt man die geschätzten Itemschwierigkeiten in einer Teilpopulation gegen die entsprechenden Schätzwerte in der Gesamtpopulation oder einer anderen Teilpopulation auf, so fallen Items mit merklichen DIF-Werten sofort auf (vgl. Abb. 3 und 4).

Bei der Beurteilung von Effekten, wie sich in den folgenden Abbildungen zeigen, sollte man zwischen Aussagen „der Effekt einer Größe ist signifikant bzw. nicht signifikant“, und „der Effekt einer Größe ist substantiell bzw. nicht substantiell“ unterscheiden. Signifikant sind wegen des großen Stichprobenumfangs von etwa 34000 Probanden Effekte sehr schnell. Wir haben bei einem Item dann von einem „substanziellem DIF“ bzgl. Zweier betrachteter Populationen gesprochen, wenn die Differenz der Itemschwierigkeiten mindestens 0,50 *logits* beträgt, das sind etwa 7 Punkte auf der internationalen Skala. Der Streifen mit der gestrichelten Berandung enthält dann die Items ohne *substanziellen* DIF.

4. Ergebnisse aus den Bildungsgängen

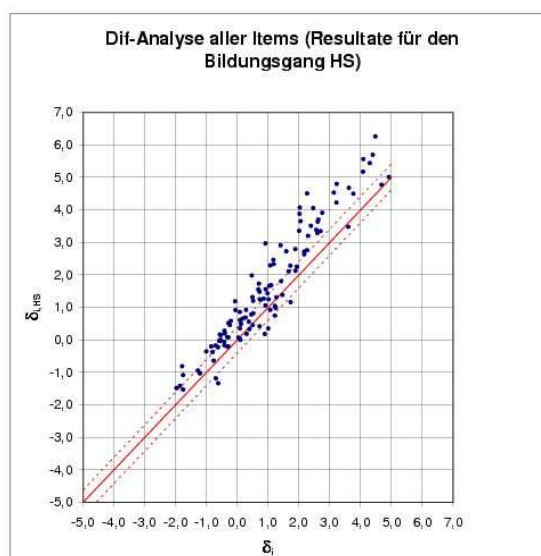


Abbildung 3

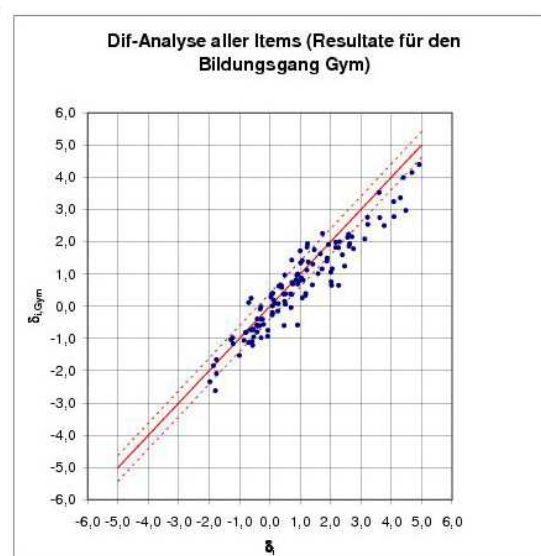


Abbildung 4

Die Abbildungen zeigen, dass sich in den „Extrempopulationen“ *Hauptschule (HS)* und *Gymnasium (Gym)* für eine große Zahl von Items substanzieller DIF nachweisen lässt. In der HS fallen DIF-Aufgaben in der Mehrzahl spezifisch schwerer gegenüber der Gesamtpopulation, im Gymnasium ist es erwartungsgemäß umgekehrt. Man kann das nicht nur mit Bedingungen erklären, die nur mit dem *Leistungsgefälle* zwischen den Schülerpopulationen gekoppelt sind. Dass weitere, insbesondere curriculare Besonderheiten verantwortlich sein müssen, zeigt ein Vergleich der HS mit der Integrierten Gesamtschule (IGS) in KNOCHE, N., LIND, D. (2004b).

Obwohl die HS und die IGS nach ihrem mittleren Leistungsvermögen nur 0,08 *logits* auseinander liegen, das sind rund 5 Punkte auf der internationalen Skala, findet man in der IGS lediglich 15 Items mit einem substanziellen DIF-Wert gegenüber 81 in der HS. Von diesen 15 Items fallen dann aber in der IGS 14 Items leichter als in der Gesamtpopulation. Für Gymnasiasten haben zwar fast alle der 54 DIF-Aufgaben, d.h. Aufgaben, die einen DIF aufweisen, einen geringeren Schwierigkeitsgrad als in der Gesamtpopulation, es gibt aber auch 8 Items, die dieser Probandengruppe schwerer fallen. In der Hauptschule fallen fast alle der 81 DIF-Aufgaben schwerer als in der Gesamtpopulation, aber 6 Items fallen Hauptschülern leichter. Vergleicht man nur Hauptschule und Gymnasium, so zeigt sich überraschend, dass 15 der Items den Hauptschülern substanziell *leichter* fallen als den Gymnasiasten

Was macht die unterschiedlichen Schwierigkeiten zwischen den Bildungsgängen aus?

Mit der Beantwortung dieser Frage betritt man das Feld der Analyse von Schwierigkeit bedingenden Faktoren auf der Grundlage von Aufgabenmerkmalen. Wir sind der Frage mit Blick auf die Kategorien *Stoffgebiete*, *Typen mathematischen Arbeitens* und *Itemformate* nachgegangen.

Der angesprochene Einfluss des Curriculums auf die Itemschwierigkeiten zeigt sich besonders bei Betrachtung der Kategorie *Stoffgebiete*. Bei den bereits erwähnten 8 Aufgaben mit positivem DIF im Bildungsgang *Gymnasium*, die also für Gymnasiasten einen höheren Schwierigkeitsgrad haben als in der Gesamtpopulation, handelt es sich bei drei Aufgaben um Aufgaben aus dem Gebiet *Umgehen mit Daten* und bei zwei Aufgaben um Items zur *Stochastik*. Diese Gruppe ist also hier bedeutend überrepräsentiert. In der Realschule und der IGS ist keine der Aufgaben aus diesen Gebieten mit einem positiven DIF versehen.

In der *Hauptschule* sind die Aufgaben zur Arithmetik und Algebra zusammengefasst überrepräsentiert. Fast 90% der Aufgaben des Gesamttests (29 von 32) aus diesem zusammengefassten Stoffgebiet erhalten einen positiven DIF-Wert, d.h. fallen Hauptschülern deutlich schwerer. Darunter befinden sich alle 21 Testaufgaben aus dem Gebiet Algebra.

In der IGS und der RS liegen die DIF-Werte generell nur geringfügig über dem Niveau, das wir mit substantiell bezeichnet haben.

Bei den 15 Aufgaben, die Gymnasiasten schwerer fallen als Hauptschülern, handelt es sich bei 13 Aufgaben um MC-Aufgaben im mittleren Schwierigkeitsbereich. Hier scheint das *Itemformat* eine Rolle zu spielen.

Was die Kategorie *Typ mathematischen Arbeitens* betrifft, lassen sich für den Bildungsgang *Realschule* wegen der geringen Zahl von DIF-Items keine Aussagen aus der Verteilung herleiten. Ansonsten verteilen sich in den einzelnen Bildungsgängen die DIF-Items auf die Kategorien in etwa so, wie es dem Anteil ihrer Kategorie im Test entspricht. Erkennbar ist ein deutlicher Unterschied zwischen den *Profilen* der HS und der IGS.

5. Einflüsse der Variablen Geschlecht

Sowohl in der Gesamtpopulation wie in den einzelnen Bildungsgängen zeigen sich Unterschiede in den durchschnittlichen Leistungsfähigkeiten von Jungen und Mädchen. Mit Blick auf die Leistungsbandbreite in der Gesamtpopulation wie in den einzelnen Bildungsgängen sind diese Unterschiede aber marginal, was wir bereits in dokumentiert haben. Das zeigt auch die Abbildung, in der die Itemschwierigkeiten für die Mädchen gegen die der Jungen aufgetragen sind (vgl. M. NEUBRAND, Hrsg, 2004, S. 81).

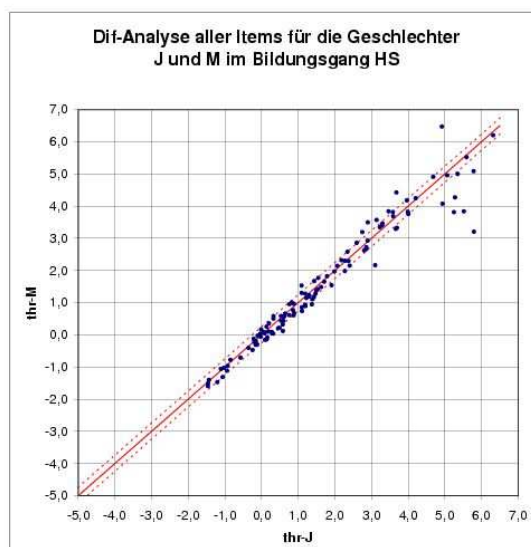


Abbildung 5

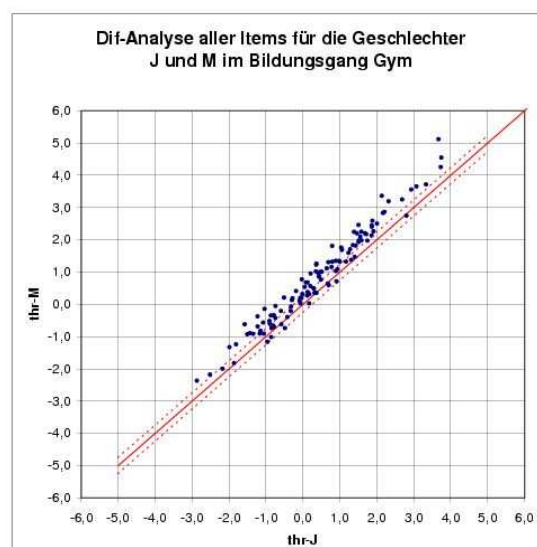


Abbildung 6

Schaut man wie in Abb. 5 und Abb. 6 in die Bildungsgänge, so erkennt man Feinheiten. Das auffälligste Ergebnis ist, dass sich das Bild mit wachsendem Bildungsanspruch zugunsten der Jungen verschiebt. Das heißt: Während es in der Hauptschule eine Teilpopulation von Schülerinnen gibt, die bei einigen der dort schwierigeren Mathematikaufgaben ihren männlichen Mitschülern überlegen sind, ist dies mit wachsendem Anspruch der Bildungsgänge immer weniger der Fall. Im Gymnasium fallen den Mädchen alle DIF-Aufgaben schwerer, wenn auch zum Teil nur geringfügig.

Bei den 6 Items, die Hauptschülerinnen leichter als Hauptschülern fallen, fällt auf, dass es sich mit um die schwierigsten Items handelt. Keines dieser Items hat das Format MC und 5 Items gehören der Kategorie *begriffliches Modellieren* an. Da alle 6 Aufgaben als *multipl lösbar* gelten und vier dem Stoffgebiet *Algebra* angehören, kann man sie dem Anspruch nach als „gymnasial“ einstufen. Hier könnte die Vermutung aufscheinen, dass es sich bei der Beobachtung um einen (*Nicht*)*selektionseffekt* handelt. In der Hauptschule gibt es wohl Mädchen, denen trotz ihrer Fähigkeiten kein adäquater Bildungsgang nahe gelegt wurde.

Literatur

- Artelt, C., Demmrich, A., Baumert, J. (2001): Selbstreguliertes Lernen. In Baumert, J. et al., PISA 2000 - Basiskompetenzen von Schülerinnen und Schülern im internationalen Vergleich. Leske + Budrich, Opladen 2001, 271-298
- Knoche, N., Lind, D. et al. (2002): Die PISA 2000- Studie , einige Ergebnisse und Analysen. *Journal für Mathematikdidaktik*, 23 (3/4), 159-202
- Knoche, N., Lind, D. (2004a): Bedingungsanalysen mathematischer Leistung: Leistungen in den anderen Domänen, Interesse, Selbstkonzept und Computernutzung. In M. Neubrand, Hrsg., Mathematische Kompetenzen von Schülerinnen und Schülern in Deutschland, Vertiefende Analysen im Rahmen von PISA 2000, VS Verlag für Sozialwissenschaften/GWV Fachverlage, Wiesbaden 2004, 205-226
- Knoche, N., Lind, D. (2004b): Eine differenzielle Itemanalyse zu den Faktoren Bildungsgang und Geschlecht. N: Neubrand, M. (Hrsg.), Mathematische Kompetenzen von Schülerinnen und Schülern in Deutschland, Vertiefende Analysen im Rahmen von PISA 2000, VS Verlag für Sozialwissenschaften/GWV Fachverlage, Wiesbaden 2004, 73-86